American Journal of Obstetrics and Gynecology (2006) ■, ■–■

ELSEVIER

American Journal of
Obstetrics & Gynecology

www.ajog.org

Genetic epidemiological studies of preterm birth: Guidelines for research

Craig E. Pennell, MD, PhD,[a,b,*] Bo Jacobsson, MD, PhD,[c,d] Scott M. Williams, PhD,[e] Rebecca M. Buus, PhD,[f] Louis J. Muglia, MD, PhD,[g] Siobhan M. Dolan, MD, MPH,[h] Nils-Halvdan Morken, MD,[i] Hilmi Ozcelik, PhD,[j] Stephen J. Lye, PhD,[k] PREBIC Genetics Working Group,[k] Caroline Relton, PhD[l]

School of Women's and Infants' Health, The University of Western Australia,[a] Perth, Western Australia, Australia; Department of Obstetrics and Gynecology, University of Toronto,[b] Toronto, Canada; Perinatal Center, Department of Obstetrics and Gynecology, Sahlgrenska University Hospital,[c] Göteborg, Sweden; North Atlantic Neuro-Epidemiology Alliances, Department of Epidemiology and Social Medicine, University of Aarhus,[d] Aarhus, Denmark; Department of Medicine, Center for Human Genetics Research, Vanderbilt University,[e] Nashville, TN; Division of Reproductive Health, Maternal and Infant Health Branch, Centers for Disease Control and Prevention,[f] Atlanta, GA; Division of Pediatric Endocrinology and Diabetes, Washington University School of Medicine,[g] St. Louis, MO; Department of Obstetrics and Gynecology and Women's Health, Albert Einstein College of Medicine,[h] Bronx, NY; Department of Obstetrics and Gynecology, Telemark Hospital,[i] Skien, Norway; Samuel Lunenfeld Research Institute, Mount Sinai Hospital,[j] Toronto, Canada; The International Preterm Birth Collaborative (PREBIC), Genetics Working Group,[k] (www.prebic.org); School of Clinical Medical Sciences (Child Health), University of Newcastle,[l] Newcastle upon Tyne, United Kingdom

Received for publication August 5, 2005; revised February 25, 2006; accepted March 13, 2006

KEY WORDS
Preterm premature rupture of membranes
Phenotype
Genotype
Genetic association study
Study design
Candidate gene
Control population
Single nucleotide polymorphism

Over the last decade, it has become increasingly apparent that the etiology of preterm birth is multifactorial, involving both genetic and environmental factors. With the development of new technologies capable of probing the genome, exciting possibilities now present themselves to gain new insight into the mechanisms leading to preterm birth. This review aimed to develop research guidelines for the conduct of genetic epidemiology studies of preterm birth with the expectation that this will ultimately facilitate the comparison of data sets between study cohorts, both nationally and internationally. Specifically the 4 areas addressed in this review included: (1) phenotypic criteria, (2) study design, (3) considerations in the selection of control populations, and (4) candidate gene selection. This paper is the product of discussions initiated by the authors at the 3rd International Workshop on Biomarkers and Preterm Birth (PREBIC) held at the University of California, Los Angeles, Los Angeles, California, in March 2005.

Supported by The University of Western Australia; the Thrasher Research Foundation; the March of Dimes Birth Defects Foundation, White Plains, NY; and the Westlakes Research Institute.

* Reprint requests: Craig Pennell, MD, PhD, School of Women's and Infants' Health, The University of Western Australia, King Edward Memorial Hospital, 374 Bagot Road, Subiaco, Western Australia 6008, Australia.

E-mail: craig.pennell@uwa.edu.aw

0002-9378/$ - see front matter © 2006 Mosby, Inc. All rights reserved.
doi:10.1016/j.ajog.2006.03.055

Preterm birth (PTB), birth at less than 37 completed weeks of gestation, is the most significant clinical problem facing contemporary perinatology in the developed world. It is the single most important cause of perinatal mortality and morbidity in industrialized countries: 60% to 80% of deaths of infants without congenital anomalies are related to preterm birth.[1] Furthermore, PTB is associated with cerebral palsy and other long-term health sequelae including cognitive impairment, blindness, deafness, respiratory illness, and complications of neonatal intensive care.[2-4]

Despite major advances in our understanding of both term and preterm labor, over the past 2 decades, the rate of PTB has been escalating steadily and alarmingly. In the United States between 1981 and 2002, the rate of PTB rose 29%, from 9.4% to 12.1%.[5-7] This increase has resulted in more than 470,000 babies being born preterm every year in the United States.[5,6] Similar increases in the incidence of PTB have been reported in Canada,[8] Australia,[9] and Denmark.[10] The only developed countries to report a decrease in PTB over the same time period are France, Finland, and most recently Sweden.[11,12]

Over the last decade, it has become increasingly apparent that the cause of PTB and preterm premature rupture of the membranes is multifactorial and involves both genetic and environmental factors.[13-18] Similar observations have been made in other complex diseases such as coronary heart disease,[19,20] hypertension,[21,22] depression,[23,24] and other psychiatric conditions.[24] Preterm delivery has been shown to be familial in a study of the heritability of PTB which demonstrated that probands with PTB were more closely related to each other than random members of the population.[25] Furthermore, twin studies have suggested that heritability for PTB ranges from 17% to 36%.[26,27] Transgenerational studies have also demonstrated an increased risk of PTB for women who themselves were born preterm.[28]

The risk of PTB increases as the gestational age of the mothers' birth decreases with mothers born at less than 30 weeks having a 2.4-fold (95% confidence interval 1.4 to 4.2) increase in risk of PTB.[28] Furthermore, the well-described racial differences in the rate of PTB (even when controlling for other contributing etiological factors[29]) suggests the contribution of genetic factors. The single best predictor of PTB among multiparous pregnant women is a past history of preterm delivery: women with 1 prior PTB have a recurrence risk of PTB of 15% and those with 2 prior PTBs have a recurrence risks of 32%.[30] The risk of PTB tends to remain with the mother through multiple pregnancies, even with increased levels of prenatal surveillance and preventive interventions. This finding and the coupled observations of a familial tendency for PTB[29] and racial differences in PTB rates suggest that there may be genetic components to the risk of PTB.

With the disclosure of the sequence of the entire human genome[31,32] and the availability of high-throughput methods making genotyping of large numbers of samples faster and less expensive than ever, our ability to acquire genetic data has increased exponentially. These technological advances now offer exciting possibilities to gain an entirely new insight into the mechanisms leading to PTB and has prompted a host of association studies investigating the relationship between specific polymorphic variants and various aspects of the preterm birth phenotype (see review[18]). Many of these studies have been limited by small study size, publication and reporting bias, improper study design, and lack of common standards worldwide. In order for researchers in PTB to participate in and benefit from international efforts such as the Network of Networks[33] (which aims to offer methodological support, promote sound study design, and promote standardized analytical practices for genetic epidemiological studies), guidelines for research are needed.

This paper aimed to develop research guidelines for the conduct of genetic epidemiological studies of preterm birth. It is anticipated that this will ultimately facilitate the comparison of data sets between study cohorts both nationally and internationally. This paper is the product of discussions by the authors at the 3rd International Workshop on Biomarkers and Preterm Birth (PREBIC), which was held at the University of California, Los Angeles, Los Angeles, California, in March 2005. Guidelines for the design, execution, and interpretation of genetic association studies designed to identify genetic determinants of susceptibility to complex diseases have been detailed elsewhere and provide an excellent reference point.[34,35] Specifically, issues such as population stratification, data-driven subgroup analysis, possible absence of linkage disequilibrium between marker and disease locus, and testing of multiple hypotheses have been reviewed by Keavney[34] and issues related to the application of genetic association studies to the field of reproductive biology have been reviewed by Romero et al.[35]

In this paper we will consider genetic association studies specifically in the context of PTB. The 4 areas to be addressed include: (1) phenotypic criteria, (2) study design, (3) considerations in the selection of a control population, and (4) candidate gene selection.

Phenotypic criteria

The first essential requirement to conduct genetic association studies to decipher a complex disease such as preterm birth is the development of a standardized definition of outcomes: the phenotype. Preterm birth, as defined by the World Health Organization, is birth before 37 weeks' gestational age or before 259 days.[36]

In this definition, the lower limit is not specified. To gain meaningful information from genetic association studies, we need to go beyond this classical definition because of the etiological heterogeneity of PTB, emphasizing its multifactorial origins.[37-40]

After excluding multifetal pregnancy, malformations and intrauterine fetal death, preterm birth results from 3 broad clinical conditions: (1) preterm labor leading to PTB (idiopathic PTB), (2) preterm premature rupture of membranes (PPROM), and (3) medically indicated (iatrogenic) PTB. Preterm labor leading to PTB and PPROM are often grouped together and called spontaneous preterm birth[12] because in both cases the initiation of labor is spontaneous. This is in contrast to indicated preterm birth, in which the decision to induce labor or perform a cesarean section at less than 37 weeks' gestation is iatrogenic. Preterm labor leading to PTB accounts for approximately 50% of all PTB (range 23.2% to 64.1%).[37,41] It is more frequent in populations without any established risk factors in which it represents up to 50% to 70% of all preterm deliveries according to the populations studied.[42,43]

A number of risk factors have been reported for preterm labor leading to PTB including: personal history of previous PTB, low body mass or poor weight gain during pregnancy, obesity, strenuous physical workload or ergonomic factors, uterine anomalies, psychological stress, smoking, drug abuse, in vitro fertilization, and extremes of maternal age (less than 18 years or 40 years of age or greater).[30,37,44-46] PPROM, which is usually followed by preterm delivery within 2 to 7 days, accounts for another 25% of all preterm births (range 7.1% to 51.2%).[37] Infection is usually regarded as the main cause of PPROM, and it occurs more commonly among women of low socioeconomic status and among black women.[41] Medically indicated PTB (in the absence of PPROM or preterm labor leading to PTB) occurs in about 25% of all PTB with variations from 8.7% to 35.2% according to studied populations.[37,47] Medical indications relate to both compromise in fetal well-being such as being small for gestational age or nonreassuring fetal status and maternal complications such as severe pre-eclampsia or antepartum hemorrhage.

The contribution of each of the clinical groupings that result in preterm birth varies across gestation (Table I); however, the proportion of spontaneous PTB and medically indicated PTB is relatively consistent between populations (especially in the series published after 1986), even though the rate of PTB varies up to 3-fold between populations (Table II).

There is currently no consensus about whether to aggregate or disaggregate PTB from these 3 clinical conditions in studies of etiology of PTB. A key argument for grouping all PTB is that the conditions that motivate medical intervention for early delivery (eg, preeclampsia, small for gestational age) share mechanisms such as inflammation and vascular compromise with the pathways that lead to spontaneous PTB.[48-50] If the etiologies are indeed shared, grouping together offers increased statistical power in the study of determinants.[48]

The alternate argument is to split preterm births into subsets that arise from the diverse clinical pathways that can lead to PTB. Spontaneous PTB is clinically quite distinct from severe pre-eclampsia or being small for gestational age that needs to be managed by medically indicated PTB. Although splitting all preterm births (a heterogeneous group) into subsets will result in smaller numbers of patients in each group, the increase in homogeneity in the study groups may offer increased sensitivity to detect differences in genetic epidemiology studies of preterm birth. An alternative to subgroup analyses are covariate-based analyses that can incorporate quantitative traits directly. This approach may be a more appropriate framework for the complex situation of PTB.

The PREBIC genetics working group believes that although grouping all types of PTB might be appropriate for evaluating demographic or clinical associations with PTB, for genetic studies we would strongly advocate, where possible, to subset PTB into preterm labor leading to PTB (idiopathic PTB), PPROM, and medically indicated PTB. Few studies evaluating risk factors for PTB have empirically evaluated the effect of grouping all PTB together versus analyses based on subsets determined by clinical data. Savitz et al[48] recently reported a direct comparison of spontaneous PTB (idiopathic PTB and PPROM) and medically indicated PTB for demographic and clinical predictors of PTB. Although the influences of many risk factors were shared across the 2 groupings, inconsistent or divergent relationships were identified for a number of risk factors between the groupings. This is not surprising, given the complexity of the etiological pathways in PTB.

To date, although a number of studies have been published evaluating associations between specific polymorphisms and spontaneous preterm birth,[51,52] no genetic epidemiology studies have empirically evaluated the effect of grouping all PTB together versus subsetting them into preterm labor leading to PTB (idiopathic PTB), PPROM, and medically indicated PTB, most likely because of small subgroup sizes precluding informative statistical analyses. A further argument for separating PTB into clinically determined subgroups is that these subgroups may be associated with varying antepartum treatment protocols (eg, antibiotics, corticosteroids, tocolytic therapy, or antihypertensive prior to delivery) and different long-term neonatal outcomes. For example, there are data suggesting an association between spontaneous PTB and cerebral palsy,[53,54] whereas medically indicated PTB appears to be associated with neonatal respiratory distress syndrome, retinopathy of prematurity, and broncopulmonary dysplasia.[54]

Table I Classification of preterm birth by gestational age

Subgroups	Less than 28 wk, %	28-31 wk, %	32-33 wk, %	34-36 wk, %	Less than 37 wk, %
Spontaneous preterm birth	49.5	35.6	42.6	60.6	55.2
Iatrogenic preterm birth	17.4	26.7	23.9	18.7	20.2
Intrauterine fetal death	2.3	9.0	4.6	1.4	2.7
Malformations	4.7	5.5	5.6	4.3	4.6
Multiple birth	16.0	14.7	16.0	10.1	11.6
Unknown onset of delivery	10.1	8.5	7.3	4.9	5.7
Total	100	100	100	100	100

Data from Swedish population birth statistics 1991 to 2001 in which there were 1.2 million births[12] (reprinted with permission from Morken NH, Källen K, Hagberg H, Jacobsson B. Preterm birth in Sweden 1973-2001: rate, subgroups and effect of changing patterns in multiple births, maternal age and smoking. Acta Obstet Gynecol Scand 2005;84:558-65).

Table II Comparison of spontaneous preterm birth and medically indicated preterm birth between populations

Author	Year	Preterm birth rate, %	Country	Population	Medically indicated preterm birth, %	Spontaneous preterm birth, %*
Arias and Tomich[128]	1982	8.6	United States		35.2	64.8
Main et al[129]	1985	15.4	United States	Black	24.1	75.8
Piekkala et al[130]	1986	6.6	Finland	White, regional population based	28.8	71.2
Meis et al[131,†]	1987	5.7	United States	White	18.7	81.3
Meis et al[132]	1987	6.1	United States	Black, private clinic	20.0	80.0
Meis et al[132]	1987	10.8	United States	Black, public clinic	20.9	79.1
Morken et al[12]	1991-2001	5.6	Sweden	Mainly white, nation-based setting	20.3	80.7
Zhang et al[133]	1992	16.7	United States	Black	15.0	85.0
Zhang et al[133]	1992	8.0	United States	White	17.4	82.6

* Spontaneous PTB is idiopathic PTB and PPROM leading to spontaneous PTB.

† Preterm birth defined as birth weight less than 2500 g.

The PREBIC genetics working group believes that it is necessary to define the minimum phenotype information that should be reported for genetic epidemiology studies of PTB. To achieve this, we propose both a minimum data set (Table III) and an optimal data set (Table IV) for these studies. The descriptors in the minimal data set have been selected to encompass variables that will allow specific subgroups of PTB to be identified and include variables with strong associations with increased risk of PTB. The more expansive optimal data set contains a larger additional set of variables that will allow further subgrouping of women with PTB and again includes variables associated with increased risk of PTB.

The idea of having a defined minimum standard for information associated with experiments is not new to life sciences. Similar minimum data sets have been described in other areas of genomic research such as the minimum information about a microarray experiment (MIAME) guidelines for microarray studies that have been developed by the Microarray Gene Expression Data Society[55] and the mode of operations adopted by the macromolecular structure community (for example, http://msd.ebi.ac.uk/), in which most journals require submission of a well-defined minimum of raw data associated with publications.

Use of the minimum and/or optimal data sets described in this paper for genetic epidemiology studies into PTB will provide the opportunity for the following: (1) precise interpretation of experimental results, (2) potential independent verification in different populations, (3) comparison of results between multiple studies, and (4) combining data sets from different studies without requiring further data abstraction from clinical records.

Study design

A variety of observational epidemiological approaches have evolved to permit the assessment of contribution of genetic factors to disease risk. The advantages and

Table III Minimum data set for genetic epidemiology studies into preterm birth

Minimum data set

- Spontaneous initiation of preterm birth*
  - Preterm labor leading to PTB (idiopathic PTB)
  - PPROM†
- Medically indicated preterm birth (nonspontaneous initiation)
- Living fetus versus intrauterine fetal death when commences labor
- Singleton or multifetal pregnancy
- Gestational age at delivery (utilizing American College of Obstetrics and Gynecology guidelines for dating[64])
  - Birth between 20 and 37 weeks of gestation
- Smoking status during pregnancy‡
- Use of drugs (nonprescription) and/or alcohol during pregnancy§
- Maternal variables
  - Age
  - History of previous preterm birth
  - Parity
- Ethnicity

* Preterm birth is birth between 20 and 37 weeks gestational age.

† PPROM is spontaneous rupture of the membranes prior to the onset of labor and before 37 weeks' gestation.

‡ Any smoking during pregnancy; ideally the amount, duration, and gestational period of exposure should be recorded; however, frequently little sensitivity or specificity is gained beyond a yes/no response because most women will underreport their smoking habits.

§ Any use of nonprescription drugs or alcohol during pregnancy: type of drug, frequency, and gestational period of exposure.

Table IV Optimal data set for genetic epidemiology studies into preterm birth

Optimal data set*

- Demographic variables
  - Type of prenatal care
  - Socioeconomic status
  - Maternal education
- Clinical variables
  - Spontaneous labor versus induction of labor
- Maternal variables
  - Height/prepregnancy weight/body mass index
  - Maternal nutritional status
  - Weight gain during pregnancy
  - Uterine anomaly
  - Psychological stress
  - Use of medication during pregnancy
    1. Tocolytic therapy (timing and duration)
       - i. Prior to clinical presentation (eg, progesterone)
       - ii. At the time of clinical presentation
    2. Antibiotics
  - Previous cervical conization/loop electrosurgical excision procedures
  - Cervical cerclage
  - Mode of conception
    1. Natural
    2. Assisted (ovulation induction, in vitro fertilization, intracytoplasmic sperm injection, donor sperm, donor egg, etc)
  - Evidence of infection
    1. Fever
    2. Tachycardia
    3. Placental histopathology
  - Pre-existing medical conditions
    1. Hypertension
    2. Diabetes
    3. Autoimmune conditions
  - Complications of pregnancy
    1. Pre-eclampsia/eclampsia
    2. Abruption
    3. Recurrent antepartum hemorrhage
    4. Small for gestational age
- Fetal variables
  - Birth weight
  - Congenital anomaly
  - Evidence of infection
    1. Tachycardia
    2. Amniotic fluid assessment
    3. Early neonatal infection
- Placental histopathology
  - Infection
  - Uteroplacental ischemia
- Family history
  - Maternal gestational age at delivery
  - Mother, father, or sibling with history of PTB
    1. Validated or self-report

* Not in order of importance.

disadvantages of each of these approaches are summarized in the third of a very informative series of 7 articles on genetic epidemiology.[56]

Cohort studies are considered the most robust of all observational designs because many issues relating to bias and confounding can be overcome by adopting this study design. One firm advantage of adopting a cohort study design is the facility to consider gestational age as a continuous variable rather than introducing an arbitrary 37-week cut-off point; however, the incidence of PTB entails large sample sizes in these studies, which in turn inevitably entails large costs.

The most commonly adopted alternative is a case-control design, in which affected and unaffected individuals are sampled. These tend to be simpler, easier to administer, and more cost efficient, hence their popularity. In practice a nested case-control study set within a prospectively collected cohort might be the preferred design to avoid selection bias (see section on selection of a control population).

The design of studies to identify genetic risk factors for PTB presents all of the usual problems for genetic studies of complex disease that have been reviewed extensively elsewhere and will not be discussed in detail here (see recent review[57]); however, in addition to these issues, PTB presents unique problems. Specifically, it is possible that the maternal genotype predisposes to PTB, the fetal genotype predisposes to PTB, or both interact to do so. If one considers that genetic risks are not causative in the same way that mutations cause Mendelian disease, but affect only risk or susceptibility, this creates a set of analytical issues that make the design of studies and the interpretation, if not the collection of data, especially difficult. Study design needs to reflect this added complexity.

To appropriately design genetic studies, we must first consider who is the affected individual or case. In more traditional disease studies this is straightforward: an individual either has the disease or does not; however, in PTB it is less clear cut who is affected. Is it the mother or the fetus? In one sense it is neither but more precisely the birth event. Using this as the conceptual framework allows a variety of designs to be used, but are any of them really optimal or do we need to take a more pluralistic approach?

The potential for interaction among different genetic factors in different individuals has not been assessed extensively, although a few methods have been developed to study this possibility.[58-62] These papers attempt to use a variety of related statistical techniques to address the effects of both maternal and fetal genotype. The most commonly used example of this is the transmission disequilibrium test to test for linkage disequilibrium between a marker and a putative disease locus using case-parent trios. Such approaches have the advantage of overcoming population stratification by using within family controls and allow the possibility of using expectation maximization algorithms to compensate for a certain degree of missing data within each trio.[62] Other alternatives include the method of Wilcox et al,[61] which assess the effect of fetal genotype on PTB using a log linear model testing for an overrepresentation of inheritance of specific alleles in preterm events. In the case of the mother's role, the test assesses the differences between the mother's and father's genotypes. This approach allows for the assessment of maternal contribution, the fetal contribution and both to the pregnancy outcome.

An alternative approach to the case-only, family-based design summarized earlier is to use a more classical case-control approach. In this scenario both preterm and term birth events are included in the analysis, and allele and/or genotype frequency differences are assessed in PTB versus term births. This approach is a standard genetic epidemiological approach and has the logical advantage that it is not necessary to recruit fathers into the study, a process not always possible; however, the question of whose genotype is critical is still an issue.

Several analytical approaches are possible to address this issue. One is to assess an association between maternal and fetal genotypes independently. This, however, ignores the aforementioned possibility of interaction. An alternative to separate analyses is to use statistical approaches such as logistic regression with both genotypes as variables. Although this is a possible approach, it does have some caveats; because the maternal and fetal genotypes are not independent of each other, they are not truly uncorrelated variables and this makes analyses more difficult. This issue will require more attention by statistical geneticists in the near future. At present most investigators have analyzed maternal and fetal separately and drawn conclusions about the contribution of each.[14,16,63-71] In the long term, this will not be adequate.

Furthermore, in addition to potential maternal-fetal genotype interactions, there is the distinct possibility that multiple genes act together to affect pregnancy outcomes. This concept has begun to get increasing attention in the genetic analysis of other complex disease.[72-75] Gene-gene interaction (or epistasis) can operate in several ways.[74] First, the effect of 1 gene may be masked by the genotype at a second locus, making it difficult to infer a genotype-phenotype relationship without knowledge about the second locus. Second, each genotype may have distinct phenotype, but the phenotype will differ as a function of genetic background. Third, effects at 2 loci may be present and detectable, but the effect of the 2 loci may be additive or multiplicative.

It should be noted that these types of contextual effects of genetic background can also occur with environmental context such that the genetic effects are distinct across environments (gene-environment interactions). Failure to consider such context can cause research results to be inconsistent across studies, a not uncommon situation in the genetic study of complex disease. The net effect of these types of interactions will be to maintain homeostasis (eg, normal gestational length) in the presence of perturbations by compensating at other genes for deficiencies in 1 gene.[76] Only when variants at more than 1 gene prevent such compensation will a disease phenotype present.

Several approaches have been developed that can handle gene-gene or gene-environment interaction, although none of them is ideal in all situations.[77] Most of them are best suited to a case-control design as opposed to family-based approaches; however, as with the other designs to study of the genetics of PTB, it is unclear how to include both maternal and fetal genetic information. These approaches include the following: (1) logistic regression and stepwise regression[78]; (2) set association[79]; (3) classification and regression trees; (4) multivariate adaptive regression spline[80]; (5) focused interaction testing framework[81]; and (6) neural networks[82] for samples of unrelated individuals. Each of these approaches has limitations that include assumptions with regard to single-locus effects, assumptions with regard to either a statistical or genetic model, and the so-called curse of dimensionality, the state of data being too sparse for the number of loci being studied to estimate genetic effects: yet each of these approaches is more powerful in certain situations than the others. A discussion of many of these methods is outside the scope of this paper but can be can be found elsewhere.[72]

Another approach that has been used with increasing frequency is multifactor dimensionality reduction (MDR).[83,84] This technique is effective at dealing with large numbers of potential interactions between genes and has already proven informative in studies of preterm birth.[85] MDR has the advantage that it does not assume a mode of inheritance for the disease and can identify genetic risk models at multiple loci in the absence of single-locus effects. In fact, it is more powerful than some of the methods mentioned earlier when no single-locus effects exists, although it is less powerful when single-locus effects exist.[81] The method is also tolerant of moderate levels of genotyping errors and/or missing data. MDR also incorporates a cross-validation method, not usually used in more standard analyses. Such model validation has proven to be informative in the ability to generalize results and prevent type I error.[86] MDR can, however, be computationally costly, especially testing large data sets for higher order (more than 4 loci) interactions and results may be difficult to interpret from a physiological point of view. It also has low power in cases of genetic heterogeneity and phenocopies, but these limitations are probably no different from other methods. Additionally, it may not be suited to assess the role of 2 genotypes that are not independent but in which both affect phenotype.

As data analysis becomes more complex, simple analytical methods may be impractical and analyses may have to proceed in several steps that sequentially simplify the data set being used. Some methods noted use this paradigm as their central technique,[79,87] whereas others have implemented techniques such as MDR in a hierarchical fashion based on subdividing the data by physiological pathways.[84] Finally, the complexity of these analyses and the number of statistical tests being performed often make interpretation of results difficult or limited.

Multiple testing is an issue that has received much attention in genetic epidemiology literature. A review by Hirschhorn et al[88] reported that of 166 putative associations that had been studied 3 or more times, less that 4% were reproduced more than 75% of the time. All statistical tests should be reported, and account should be taken when testing several loci in 1 study population to minimize false positives from multiple hypothesis testing. The most commonly used method for correcting for multiple hypothesis testing is the Bonferroni correction. An alternative strategy that has been advocated to address the inherent statistical multiple comparisons problem is to adopt a 2-stage approach.[89] Other more elaborate statistical methods such as Bayesian methods that can take into account the posterior probability of the validity of an association are now being applied. The application of such approaches will be an important step forward in reducing false-positive results.

The PREBIC genetics working group recommend that authors practice full disclosure in the presentation of their analyses in terms of how many statistical tests were performed to allow readers to judge the significance of the results for themselves.[90] The exception to this would be when studies are predominantly meant as exploratory or hypothesis generating. Furthermore, power calculations should also be an integral part of a study design to ensure that the study is sufficiently powered to address the stated hypotheses.

Finally, pooling genomic deoxyribonucleic acid (DNA) samples is an attractive strategy in both case-control and family-based studies because it allows rapid and inexpensive genotyping. A variety of technologies can be applied to detect single nucleotide polymorphisms (SNPs) using pooled DNA samples including primer extension assays and microarrays; however, certain prerequisites are essential for this approach to be successful, in particular, the accurate measurement of DNA concentration followed by the preparation of replicate pools. This approach also depends greatly on the robustness of the SNP assay because the generation of false positives can drastically influence the detection of associations. Validation of pooled results against individual genotype results is an important step in optimizing assays for pooled samples. Unless very large numbers of samples are to be processed, the effort required in optimization of genotype assays for pooled samples can outweigh the potential benefits.

To summarize, there are 3 distinct designs that can be used to assess the role of genetic risks in PTB: a family-based trio design; a cohort design; and a case-control design. The former has the advantage that several analytical methods have been developed to deal with triads of collected data. The others have the advantage that the data are easier to collect but the disadvantage that analytical tools are not completely developed to assess maternal-fetal contributions simultaneously. The PREBIC genetics working group believes that the optimal study design for genetic epidemiology studies of PTB would include at a minimum mothers and affected offspring. If possible, it would be an asset to recruit fathers and even unaffected siblings; however, the practicalities of undertaking this should not be underestimated, and the costs possibly outweigh the benefits. Several approaches are now available to infer genotype and haplotype of missing members in incomplete case-parent trios.[91]

Therefore, there is currently no perfect design, and ongoing work is required to refine analytical methods to improve our ability to assess genetic risk factors of PTB. Regardless of the design to be used, it is critical that extreme care be taken in design and collection of current data so that as new and better analytical methods are developed, the data can be useful for reanalysis.

Considerations in the selection of a control population

The acquisition of suitable control populations is a critical component of any large-scale genetic association study of the etiology of PTB.[92] Appropriate control selection, in which controls accurately represent the base population from which the cases arise, can minimize systematic differences, ensure accurate interpretation of the data, and enhance the ability for findings to be replicated. A cohort study design offers the intrinsic benefit of being predicated on prospective collection of subjects from a defined population with cases and controls arising out of this single group. The same is not true of case-control genetic epidemiology studies: in these studies, when poor control selection has occurred, disease-gene associations can be performed with controls that were selected from different populations from the cases, which can result in misinterpretation of results.

To minimize systematic bias in collection of cases and controls, acquisition of these groups from a common geographic site or group of sites is critical.[93] Obtaining controls from the same sites as cases may facilitate matching the groups for socioeconomic status and pre- and perinatal care practices, both of which could affect the etiology of the preterm births analyzed. In addition to obtaining cases and controls from common sites, care should also be taken to diminish temporal drift in gene pools by concurrently collecting cases and controls.[94] Moreover, collecting controls and cases from the same geographic site(s) and time period assists in the capability of investigating gene/polymorphism-environment interactions. As with nongenetic studies, utilizing the same sites should help control for mix of race and ethnicity within a population which helps to reduce the possibility of population stratification in allele frequencies; however, in many countries in the world, this approach would not be sufficient to avoid population stratification. Furthermore, instituting measures to avoid genetic drift and population stratification cannot be overemphasized.

One method by which to confirm that controls are ethnically related to cases is to use multiple genetic markers of ethnicity.[95-98] Testing for Hardy-Weinberg equilibrium should also be a standard of practice because it is sensitive to population stratification,[99] although deviations may also be indicative of genetic association.[100] Two additional methods have been proposed that can help correct for failure to adequately match cases and controls. One method, genomic control, proposes the collection of genetic information at several loci that are unrelated to the phenotype to test for underlying population structure and to correct for it if it occurs.[101] The second method, known as structured association, tests for underlying population structure explicitly, and then this information can be used to divide samples by genetically identified groups prior to analysis for association.[102] Genetic markers of ethnicity are not yet currently available for all ethnic groups; therefore, genome control or structured association analyses should be performed in populations in which concerns of substructure arise.

The general concept for assuring valid case-control comparisons is to match these groups for noninvolved alleles and nongenetic contributions as closely as possible. Because the etiology of human preterm birth is heterogeneous, research designed to specifically identify genetic contributions to risk will need to consider the many previously appreciated nongenetic contributors to prematurity.[35] To enrich for the genetic contribution of preterm birth, cases and controls should be matched on variables for nutritional status, adequacy of prenatal care, socioeconomic status, maternal age, body mass index/prepregnancy weight/weight gain, maternal exposure history (including smoking, medications, drug use, and environmental and social exposures), and previous medical and obstetric history. Other potential variables for matching or for use as inclusion criteria include history of prior preterm birth or a specific gestational age range at delivery.

The PREBIC genetics working group believes that control population should be drawn from uncomplicated deliveries and that caution should be exercised in excluding too many neonatal phenotypes; however, because small for gestational age may share common aspects in the etiology with PTB, the inclusion of this neonatal phenotype in the control group could disguise an association. Careful description of inclusion criteria, exclusion criteria, and attempts to correct for potential biases should be incorporated in each report of findings.

The selection of cases and controls for preterm birth provides some novel methods for assigning genetic risk and defining case and control status. For example, cases could specifically be multiparous mothers with recurrent PTB versus controls of multiparous mothers with only term birth. In this situation, because it remains unclear whether the mother or fetus is the proband, multiparous mothers with recurrent preterm birth could reflect either 1 affected individual (mother) or 2 affected siblings. In these types of family-based studies, one can make comparison of affected to nonaffected relatives or of generational transmitted versus nontransmitted alleles. Family-based studies in which siblings are used as controls have the advantage that cases and controls are derived from the same overall gene pool.[92]

Alternatively, to provide even greater evidence of a common genetic contribution among affected families, it will be particularly helpful to identify families with first-degree relatives having preterm birth in comparison with pedigrees without preterm birth.

Candidate gene selection

There is a considerable body of literature attributing polymorphisms (particularly SNPs) within genes as primary contributors of an individual's risk of disease. SNPs may occur in noncoding regions of genes, in which they may have an impact on the rate of transcription and/or messenger ribonucleic acid stability, or in coding regions, in which they have the potential to result in alterations in the protein sequence. Studies of sets of human genes have demonstrated that each gene contains several SNPs, including coding region substitutions, with a frequency of greater than 1% in the general population. About half of the coding SNPs are silent, whereas the other half are nonsynonymous SNPs, resulting in amino acid substitutions.[103,104] SNPs can alter the functional output of the genome through various mechanisms including altered protein structure, function, and interaction with other proteins. SNPs conferring disease risk tend to occur in structurally and functionally important protein domains and are most likely to affect function by either decreasing cellular protein levels or altering the structure of these molecules.[105,106] How do we identify the most relevant SNPs in candidate genes to study in the context of PTB susceptibility? The most comprehensive analysis of candidate SNPs is obtained by resequencing the entire candidate gene in patients and controls to search for disease-specific variants. This approach is expensive and limited.

The study of allelic variants is a simpler, cheaper yet powerful approach, hence its popularity. Nevertheless, searching for disease causing SNPs is a challenging task because of the complex nature of the PTB phenotype and the enormous number of SNPs that could be analyzed. The situation is further complicated by all the genes involved in the complex pathophysiology of PTB that have yet to be identified. Furthermore, a candidate gene approach precludes the use of unknown genes or genes with unknown function.

Two approaches have generally been adopted to identify candidate susceptibility genes: the functional approach and the positional approach. In the functional approach, used in many earlier studies, the gene coding for a protein with a known biochemical role is investigated. This approach has been used most widely in PTB studies of genetic susceptibility, despite the drawbacks alluded to above. The positional approach uses linkage analysis to identify regions of the genome that segregate with the disease of interest. Although generally the more successful of the 2 approaches, the positional approach has been of less utility investigating PTB because extended pedigrees of affected individuals are a prerequisite for linkage analysis and few such collections of PTB families exist.

Once candidate susceptibility genes have been selected, specific SNPs within those genes need to be identified. Bioinformatics-based strategies may be instrumental in segmenting molecular studies into a subset of SNPs with a greater likelihood of conferring to specific disease phenotypes. This approach is both essential and beneficial to prioritize SNPs of genes in biological pathways associated with complex diseases. In contrast to highly penetrant mutations, common SNPs are usually associated with less dramatic effects on protein function, which, alone, may not be destructive to the capacity of the biological process; however, inheritance of combinations of functional, commonly occurring SNPs may additively or synergistically lead to an altered function, thus to a distinct phenotype.

There are a growing number of bioinformatics and computational techniques available to identify functional SNPs that are likely to affect the folding and thus function of the encoded proteins. Bioinformatics approaches include data mining from Web-based SNP databases and other relevant sources. Automated tools are available to mine and validate SNP information from both public (eg, dbSNP,[107] HGVBase,[108] GeneSNP, and SNP500) and private databases (Celera[109] [https://myscience.appliedbiosystems.com]). The 2 largest SNP databases, dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and HGVBase (http://hgvbase.cgb.ki.se/), are general databases containing more than 2 million SNP entries, whereas GeneSNP (http://www.genome.utah.edu/genesnps/, http://lpgws.nci.nih.gov/) and SNP500 (http://bumper.nci.nih.gov/home.cfm) are relatively smaller, specialized databases. Uniqueness and specificity of the identified nonsynonymous SNPs can be evaluated using the BLAST against gene transcript tool (http://lpgws.nci.nih.gov:82/perl/blast2) and BLAST against human genome tool of NCBI,[110] as explained in Savas et al.[111]

Once nonsynonymous SNPs within candidate genes have been identified, there are a suite of computational applications to predict functionality of SNPs. A multitude of Web-based tools are available to aid this process. An example of the methodology and associated Web-based tools is outlined below:

1. Protein alignment and evolutionary conservation analysis. This strategy assesses evolutionary conservation (also known as profiling modeling), which exploits the relationship between sequence conservation across species and criticality of function to detect which amino acid substitutions are most likely

to be deleterious. The 2 most commonly used tools include SIFT[112,113] (sorting intolerant from tolerant [http://blocks.fhcrc.org/∼pauline/SIFT.html]) and PolyPhen.[114]

2. Protein structure functional prediction methods. These methods consider 3-dimensional structure effects of amino acid substitution directly and identify the impact on structure stability rather than protein function. The models include structure stability modeling,[115] FOLD-X[116,117] (http://fold-x.embl-heidelberg.de:1100/cgi-bin/main.cgi) and post-translational modification (NetPhos program[118] [http://www.cbs.dtu.dk/services/NetPhos/]).

Overall the sensitivity of this suite of prediction tools to model SNP function is 70% to 90%. It has been successfully used to systematically study the DNA repair and cell cycle SNPs[111,119] and genetic variants involved in breast cancer.[120-122]

Although much of the earlier discussion relates to functional variants in coding regions of genes (coding SNPs), predisposing gene variants may not have an obvious role in gene function but instead may occur in nonexpressed sequences such as introns and flanking regions (noncoding SNPs). For example, type 2 diabetes–predisposing variants occur in the introns of calpain 10[123,124] and far upstream of the hepatocyte nuclear factor-4α gene[125] and common variation in a region of no obvious function predisposes to Hirschsprung's disease.[126] It is therefore important to capture as much of the common variation across a candidate gene as possible.

SNPs that capture a large proportion of the common variation across a given gene can be determined from the publicly available HapMap (http://www.hapmap.org/) data. To ensure that HapMap SNPs do not give redundant information because of linkage disequilibrium, genotype information from publicly available genotyped cohorts such as the 30 white trios used to generate the HapMap data can be used as a reference. The algorithms of tagger (http://www.broad.mit.edu/mpg/tagger/) can then be applied to select an efficient, nonredundant set of tagging SNPs that capture all HapMap SNPs across the gene (± 10 kb either side) at $r^2 > 0.8$. It must be noted that not all SNPs within publicly accessible databases have been validated, and care must be taken to ensure that only validated SNPs are used.

Candidate gene studies, by their very nature, rely on the prioritization of specific genes for investigation. The choice of gene is most commonly determined by clear biological plausibility, consistent with current knowledge of the pathogenesis of the disease; however, limited knowledge of the pathophysiology of preterm birth therefore limits the choice of candidate genes. Strategies that use a fishing approach are not ideal, although recent developments in the application of whole genome approaches may be very useful in the context of preterm birth.

Genome-wide association studies, using tag SNPs are now emerging as a cost-effective alternative to large-scale candidate gene association studies. It must be recognized that the same considerations in selecting stringent phenotype criteria are prerequisites for such studies. Although technology now allows a genome-wide scan of SNPs (500,000 SNPs), current bioinformatic approaches are limited in their ability to handle data sets of this magnitude. For example, a predetermined significance value of 0.05 would generate 25,000 false-positive associations if the frequency of all known SNPs were compared in any given case and control populations. Moreover, it is unlikely that any complex disease results from a single SNP; rather a more plausible explanation is that interactions of a number of SNPs with environmental factors are responsible for significant increases in disease risk.

Analysis of interactions between multiple SNPs is complex, and tools to manage interactions among a data set of 500,000 SNPs are imperfect. A more prudent approach is therefore to combine the more traditional functional approach for candidate gene selection with a bioinformatics-based approach for SNP selection. This will create a more manageable data set for statistical analyses as well as provide enhanced cost-effectiveness. Further increases in yield can be obtained by focusing on the following: (1) SNPs that are common in the study population (minor allele frequency greater than 10%), bearing in mind that ethnic variation in allele frequencies does exist; (2) SNPs with potential functional consequences (functional SNPs) that are likely to alter the expression levels (regulatory SNPs) or folding structure of the proteins (nonsynonymous SNPs); and (3) SNPs that capture the variation across a gene and can be used to assign a haplotype.

Independent validation of results is an important step in accepting or rejecting a hypothesis and is particularly important when a significant association is elicited only after post hoc stratification of data that would otherwise have yielded negative results. Further validation may involve in vitro or in vivo functional studies.

Conclusion

Preterm birth is a global obstetric challenge with 13 million preterm deliveries annually worldwide.[127] It is likely that the cause of PTB and PPROM is multifactorial and involves both genetic and environmental factors.[14-17] Although the advent of new technologies capable of probing the genome offer exciting possibilities to gain an entirely new insight into the mechanisms leading to PTB, it is important to understand that technology alone will not resolve the complex issue of the genetic susceptibility to PTB. In this paper we have given careful consideration to the PTB phenotype, study design, candidate gene selection, and the selection of appropriate control populations required to investigate the relationship between specific polymorphic variants and various aspects of the preterm birth phenotype. It is only with careful consideration of these important issues that we will be able to truly integrate genetic information into our understanding of the mechanisms leading to PTB.

Acknowledgment

The authors acknowledge the contributions made by other members of the PREBIC Genetics Working Group to the initial discussions at the 3rd International Workshop on Biomarkers and Preterm Birth that resulted in this manuscript. Specifically we acknowledge the contributions of Dr Gunvor Ekman-Ordeberg, Dr Althea Grant, Dr Jeff Murray, and Dr Pathik Wadhwa. The authors also acknowledge that the content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Goldenberg RL. The management of preterm labor. Obstet Gynecol 2002;100:1020-37.
2. Hack M, Fanaroff AA. Outcomes of children of extremely low birthweight and gestational age in the 1990's. Early Hum Dev 1999;53:193-218.
3. Ward RM, Beachy JC. Neonatal complications following preterm birth. BJOG 2003;110(Suppl 20):8-16.
4. Marlow N, Wolke D, Bracewell MA, Samara M. Neurologic and developmental disability at six years of age after extremely preterm birth. N Engl J Med 2005;352:9-19.
5. Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Munson ML. Births: final data for 2002. Natl Vital Stat Rep 2003;52:1-113.
6. Agency for Healthcare Research and Quality. 2001 Nationwide Inpatient Sample. White Plains (NY): March of Dimes Perinatal Data Center; 2003.
7. Ananth CV, Misra DP, Demissie K, Smulian JC. Rates of preterm delivery among black women and white women in the United States over two decades: an age-period—cohort analysis. Am J Epidemiol 2001;154:657-65.
8. Health Canada. Canadian Perinatal Health Report, 2003. Ottawa: Minister of Public Works and Government Services Canada; 2003.
9. Laws PJ, Sullivan EA. Australia's Mothers and babies 2002. AIHW cat. no. PER 28. Sydney: AIHW National Perinatal Statistics Unit (Perinatal Statistics Series No. 15); 2004.
10. Langhoff-Roos J, Kesmodel U, Jacobsson B, Rasmussen S, Vogel I. Spontaneous preterm delivery in primiparous women at low risk in Denmark: population based study. BMJ 2006 [Epub ahead of print].
11. Olsen P, Laara E, Rantakallio P, Jarvelin MR, Sarpola A, Hartikainen AL. Epidemiology of preterm delivery in two birth cohorts with an interval of 20 years. Am J Epidemiol 1995;142: 1184-93.
12. Morken NH, Källen K, Hagberg H, Jacobsson B. Preterm birth in Sweden 1973-2001: rate, subgroups and effect of changing patterns in multiple births, maternal age and smoking. Acta Obstet Gynecol Scand 2005;84:558-65.
13. Romero R, Mazor M, Munoz H, Gomez R, Galasso M, Sherer DM. The preterm labor syndrome. Ann N Y Acad Sci 1994; 734:414-29.
14. Annells MF, Hart PH, Mullighan CG, Heatley SL, Robinson JS, Bardy P, et al. Interleukins-1, -4, -6, -10, tumor necrosis factor, transforming growth factor-beta, FAS, and mannose-binding protein C gene polymorphisms in Australian women: risk of preterm birth. Am J Obstet Gynecol 2004;191:2056-67.
15. Gomez R, Ghezzi F, Romero R, Munoz H, Tolosa JE, Rojas I. Premature labor and intra-amniotic infection. Clinical aspects and role of the cytokines in diagnosis and pathophysiology. Clin Perinatol 1995;22:281-342.
16. Macones GA, Parry S, Elkousy M, Clothier B, Ural SH, Strauss JF 3rd. A polymorphism in the promoter region of TNF and bacterial vaginosis: preliminary evidence of gene-environment interaction in the etiology of spontaneous preterm birth. Am J Obstet Gynecol 2004;190:1504-8; discussion 3A.
17. Wang H, Parry S, Macones G, Sammel MD, Ferrand PE, Kuivaniemi H, et al. Functionally significant SNP MMP8 promoter haplotypes and preterm premature rupture of membranes (PPROM). Hum Mol Genet 2004;13:2659-69.
18. Crider KS, Whitehead N, Buus RM. Genetic variation associated with preterm birth: A HuGE review. Genet Med 2005;7:593-604.
19. Talmud PJ, Hawe E, Miller GJ. Analysis of gene-environment interaction in coronary artery disease: lipoprotein lipase and smoking as examples. Ital Heart J 2002;3:6-9.
20. Talmud PJ, Humphries SE. Gene: environment interactions and coronary heart disease risk. World Rev Nutr Diet 2004;93:29-40.
21. Pausova Z, Tremblay J, Hamet P. Gene-environment interactions in hypertension. Curr Hypertens Rep 1999;1:42-50.
22. Hamet P, Pausova Z, Adarichev V, Adaricheva K, Tremblay J. Hypertension: genes and environment. J Hypertens 1998;16: 397-418.
23. Lesch KP. Gene-environment interaction and the genetics of depression. J Psychiatry Neurosci 2004;29:174-84.
24. Lau JY, Eley TC. Gene-environment interactions and correlations in psychiatric disorders. Curr Psychiatry Rep 2004;6:119-24.
25. Ward K, Argyle V, Meade M, Nelson L. The heritability of preterm delivery. Obstet Gynecol 2005;106:1235-9.
26. Treloar SA, Macones GA, Mitchell LE, Martin NG. Genetic influences on premature parturition in an Australian twin sample. Twin Res 2000;3:80-2.
27. Clausson B, Lichtenstein P, Cnattingius S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. BJOG 2000;107:375-81.
28. Porter TF, Fraser AM, Hunter CY, Ward RH, Varner MW. The risk of preterm birth across generations. Obstet Gynecol 1997;90: 63-7.
29. Ward K. Genetic factors in preterm birth. BJOG 2003;110(Suppl 20):117.
30. Carr-Hill RA, Hall MH. The repetition of spontaneous preterm labour. Br J Obstet Gynaecol 1985;92:921-8.
31. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
32. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001; 291:1304-51.
33. Ioannidis JP, Bernstein J, Boffetta P, Danesh J, Dolan S, Hartge P, et al. A network of investigator networks in human genome epidemiology. Am J Epidemiol 2005;162:302-4.
34. Keavney B. Genetic association studies in complex diseases. J Hum Hypertens 2000;14:361-7.

35. Romero R, Kuivaniemi H, Tromp G, Olson J. The design, execution, and interpretation of genetic association studies to decipher complex diseases. Am J Obstet Gynecol 2002;187:1299-312.
36. World Health Organization. The prevention of perinatal mortality and morbidity. WHO Technical Report Series, report 457. Geneva, Switzerland: World Health Organization; 1970.
37. Moutquin JM. Classification and heterogeneity of preterm birth. BJOG 2003;110(Suppl 20):30-3.
38. Berkowitz GS, Blackmore-Prince C, Lapinski RH, Savitz DA. Risk factors for preterm birth subtypes. Epidemiology 1998;9: 279-85.
39. Berkowitz GS, Papiernik E. Epidemiology of preterm birth. Epidemiol Rev 1993;15:414-43.
40. Papiernik E, Kaminski M. Multifactorial study of the risk of prematurity at 32 weeks of gestation. I. A study of the frequency of 30 predictive characteristics. J Perinat Med 1974;2:30-6.
41. Mattison DR, Damus K, Fiore E, Petrini J, Alter C. Preterm delivery: a public health perspective. Paediatr Perinat Epidemiol 2001;15(Suppl 2):7-16.
42. Morrison JC. Preterm birth: a puzzle worth solving. Obstet Gynecol 1990;76:5S-12S.
43. Kramer MS. Determinants of low birth weight: methodological assessment and meta-analysis. Bull World Health Organ 1987; 65:663-737.
44. Cedergren MI. Maternal morbid obesity and the risk of adverse pregnancy outcome. Obstet Gynecol 2004;103:219-24.
45. Seoud MA, Nassar AH, Usta IM, Melhem Z, Kazma A, Khalil AM. Impact of advanced maternal age on pregnancy outcome. Am J Perinatol 2002;19:1-8.
46. Jacobsson B, Ladfors L, Milsom I. Advanced maternal age and adverse perinatal outcome. Obstet Gynecol 2004;104:727-33.
47. Moutquin JM, Milot Roy V, Irion O. Preterm prevention: effectiveness of current strategies. J Soc Obstet Gynaecol Can 1996;18: 571-88.
48. Savitz DA, Dole N, Herring AH, Kaczor D, Murphy J, Siega-Riz AM, et al. Should spontaneous and medically indicated preterm births be separated for studying aetiology? Paediatr Perinat Epidemiol 2005;19:97-105.
49. Klebanoff MA, Shiono PH. Top down, bottom up and inside out: reflections on preterm birth. Paediatr Perinat Epidemiol 1995;9: 125-9.
50. Thorp JMJ. Placental vascular compromise: unifying the etiologic pathways of perinatal compromise. Curr Probl Obstet Gynecol Fertil 2001;24:197-220.
51. Engel SA, Erichsen HC, Savitz DA, Thorp J, Chanock SJ, Olshan AF. Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. Epidemiology 2005;16:469-77.
52. Engel SA, Olshan AF, Savitz DA, Thorp J, Erichsen HC, Chanock SJ. Risk of small-for-gestational age is associated with common anti-inflammatory cytokine polymorphisms. Epidemiology 2005;16:478-86.
53. Dammann O, Allred EN, Veelken N. Increased risk of spastic diplegia among very low birth weight children after preterm labor or prelabor rupture of membranes. J Pediatr 1998;132: 531-5.
54. Morken NH, Källen K, Jacobsson B. Type of onset of delivery and child outcome in a nationwide population-based study among preterm children: special focus on cerebral palsy. J Gynecol Invest 2005; Abstract for Society of Gynecological Investigations, March 2005, Los Angeles, California.
55. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365-71.
56. Cordell HJ, Clayton DG. Genetic association studies. Lancet 2005;366:1121-31.
57. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6: 95-108.
58. Mitchell LE. Differentiating between fetal and maternal genotypic effects, using the transmission test for linkage disequilibrium. Am J Hum Genet 1997;60:1006-7.
59. Sinsheimer JS, Palmer CG, Woodward JA. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. Genet Epidemiol 2003;24:1-13.
60. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62: 969-78.
61. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." Am J Epidemiol 1998;148:893-901.
62. Kistner EO, Weinberg CR. Method for using complete and incomplete trios to identify genes related to a quantitative trait. Genet Epidemiol 2004;27:33-42.
63. Hao K, Wang X, Niu T, Xu X, Li A, Chang W, et al. A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods. Hum Mol Genet 2004;13:683-91.
64. Ahrens P, Kattner E, Kohler B, Hartel C, Seidenberg J, Segerer H, Genetic Factors in Neonatology Study Group, et al. Mutations of genes involved in the innate immune system as predictors of sepsis in very low birth weight infants. Pediatr Res 2004;55: 652-6.
65. Landau R, Xie HG, Dishy V, Stein CM, Wood AJ, Emala CW, et al. β2-Adrenergic receptor genotype and preterm delivery. Am J Obstet Gynecol 2002;187:1294-8.
66. Amory JH, Adams KM, Lin MT, Hansen JA, Eschenbach DA, Hitti J. Adverse outcomes after preterm labor are associated with tumor necrosis factor-alpha polymorphism -863, but not -308, in mother-infant pairs. Am J Obstet Gynecol 2004;191: 1362-7.
67. Doh K, Sziller I, Vardhana S, Kovacs E, Papp Z, Witkin SS. Beta2-adrenergic receptor gene polymorphisms and pregnancy outcome. J Perinat Med 2004;32:413-7.
68. Genc MR, Onderdonk AB, Vardhana S, Delaney ML, Norwitz ER, Tuomala RE, MAP Study Group, et al. Polymorphism in intron 2 of the interleukin-1 receptor antagonist gene, local midtrimester cytokine response to vaginal flora, and subsequent preterm birth. Am J Obstet Gynecol 2004;191:1324-30.
69. Roberts AK, Monzon-Bordonaba F, Van Deerlin PG, Holder J, Macones GA, Morgan MA, et al. Association of polymorphism within the promoter of the tumor necrosis factor alpha gene with increased risk of preterm premature rupture of the fetal membranes. Am J Obstet Gynecol 1999;180:1297-302.
70. Simhan HN, Krohn MA, Roberts JM, Zeevi A, Caritis SN. Interleukin-6 promoter -174 polymorphism and spontaneous preterm birth. Am J Obstet Gynecol 2003;189:915-8.
71. Witkin SS, Vardhana S, Yih M, Doh K, Bongiovanni AM, Gerber S. Polymorphism in intron 2 of the fetal interleukin-1 receptor antagonist genotype influences midtrimester amniotic fluid concentrations of interleukin-1beta and interleukin-1 receptor antagonist and pregnancy outcome. Am J Obstet Gynecol 2003;189:1413-7.
72. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701-9.
73. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Ann Med 2002;34:88-95.
74. Williams SM, Haines JL, Moore JH. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? Bioessays 2004;26:170-9.

75. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56: 73-82.
76. Krege JH, Kim HS, Moyer JS, Jennette JC, Peng L, Hiller SK, et al. Angiotensin-converting enzyme gene mutations, blood pressures, and cardiovascular homeostasis. Hypertension 1997; 29:150-7.
77. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 2004;20:640-7.
78. Hosmer DW, Lemeshow S. Applied logistic regression. New York (NY): Wiley; 2000.
79. Ott J, Hoh J. Set association analysis of SNP case-control and microarray data. J Comput Biol 2003;10:569-74.
80. Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 2004;23:1439-53.
81. Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 2006;78:15-27.
82. Motsinger AA, Lee SL, Mellick G, Ritchie MD. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics 2006;7:39.
83. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138-47.
84. Williams SM, Ritchie MD, Phillips JA 3rd, Dawson E, Prince M, Dzhura E, et al. Multilocus analysis of hypertension: a hierarchical approach. Hum Hered 2004;57:28-38.
85. Menon R, Velex DR, Simhan HN, et al. Multilocus interactions at maternal TNF-alpha, IL-6 and IL-6R genes predict spontaneous preterm labor in European-American women. Am J Obstet Gynecol 2006, in press.
86. Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Gaziano JM, Ridker PM, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC Bioinformatics 2004;5:49.
87. Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, et al. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet 2000;64: 413-7.
88. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002;4:45-61.
89. Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med 2005 [Epub ahead of print].
90. Ott J. Issues in association analysis: error control in case-control association studies for disease gene discovery. Hum Hered 2004; 58:171-4.
91. Guo CY, DeStefano AL, Lunetta KL, Dupuis J, Cupples LA. Expectation maximization algorithm based haplotype relative risk (EM-HRR): test of linkage disequilibrium using incomplete case-parents trios. Hum Hered 2005;59:125-35.
92. Little J, Bradley L, Bray MS, Clyne M, Dorman J, Ellsworth DL, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol 2002; 156:300-10.
93. Cooper DN, Nussbaum RL, Krawczak M. Proposed guidelines for papers describing DNA polymorphism-disease associations. Hum Genet 2002;110:207-8.
94. Lewis SJ, Brunner EJ. Methodological problems in genetic association studies of longevity—the apolipoprotein E gene as an example. Int J Epidemiol 2004;33:962-70.
95. Devlin B, Roeder K, Wasserman L. Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. Biostatistics 2000;1:369-87.
96. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet 2000; 67:170-81.
97. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001;20:4-16.
98. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68: 466-77.
99. Ott J. Association of genetic loci: replication or not, that is the question. Neurology 2004;63:955-8.
100. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 1998;63:1531-40.
101. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997-1004.
102. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155: 945-59.
103. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:231-8.
104. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999;22: 239-47.
105. Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 1997;278: 1580-1.
106. Chakravarti A. It's raining SNPs, hallelujah? Nat Genet 1998;19: 216-7.
107. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308-11.
108. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 2002;30:387-91.
109. Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, Mayberry F, et al. The Celera Discovery System. Nucleic Acids Res 2002;30:129-36.
110. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004;32:D35-40.
111. Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. Cancer Epidemiol Biomarkers Prev 2004; 13:801-7.
112. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812-4.
113. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res 2002;12:436-46.
114. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894-900.
115. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263-70.
116. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369-87.
117. Guerois R, Serrano L. The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. J Mol Biol 2000;304:967-82.

118. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999;294:1351-62.
119. Savas S, Ahmad MF, Shariff M, Kim DY, Ozcelik H. Candidate nsSNPs that can affect the functions and interactions of cell cycle proteins. Proteins 2005;58:697-705.
120. Ozcelik H, Knight JA, Glendon G, Yazici H, Carson N, Ainsworth PJ, Ontario Cancer Genetics Network, et al. Individual and family characteristics associated with protein truncating BRCA1 and BRCA2 mutations in an Ontario population based series from the Cooperative Family Registry for Breast Cancer Studies. J Med Genet 2003;40:e91.
121. Knight JA, Onay UV, Wells S, Li H, Shi EJ, Andrulis IL, et al. Genetic variants of GPX1 and SOD2 and breast cancer risk at the Ontario site of the Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev 2004;13:146-9.
122. Figueiredo JC, Knight JA, Briollais L, Andrulis IL, Ozcelik H. Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev 2004;13:583-91.
123. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 2000;26:163-75.
124. Weedon MN, Schwarz PE, Horikawa Y, Iwasaki N, Illig T, Holle R, et al. Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. Am J Hum Genet 2003;73:1208-12.
125. Silander K, Mohlke KL, Scott LJ, Peck EC, Hollstein P, Skol AD, et al. Genetic variation near the hepatocyte nuclear factor-4 alpha gene predicts susceptibility to type 2 diabetes. Diabetes 2004;53:1141-9.
126. Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 2005;434:857-63.
127. Villar J, Abalos E, Carroli G, Giordano D, Wojdyla D, Piaggio G, World Health Organization Antenatal Care Trial Research Group, et al. Heterogeneity of perinatal outcomes in the preterm delivery syndrome. Obstet Gynecol 2004;104:78-87.
128. Arias F, Tomich P. Etiology and outcome of low birth weight and preterm infants. Obstet Gynecol 1982;60:277-81.
129. Main DM, Gabbe SG, Richardson D, Strong S. Can preterm deliveries be prevented? Am J Obstet Gynecol 1985;151:892-8.
130. Piekkala P, Kero P, Erkkola R, Sillanpaa M. Perinatal events and neonatal morbidity: an analysis of 5380 cases. Early Hum Dev 1986;13:249-68.
131. Meis PJ, Ernest JM, Moore ML, Michielutte R, Sharp PC, Buescher PA. Regional program for prevention of premature birth in northwestern North Carolina. Am J Obstet Gynecol 1987;157:550-6.
132. Meis PJ, Ernest JM, Moore ML. Causes of low birth weight births in public and private patients. Am J Obstet Gynecol 1987;156:1165-8.
133. Zhang J, Savitz DA. Preterm birth subtypes among blacks and whites. Epidemiology 1992;3:428-33.

Condensation The objective of this review was to develop research guidelines for genetic epidemiological studies of preterm birth.